Jurnal Informatika: Jurnal pengembangan IT

Vol. 11, No. 2, 2026
ISSN : 2477-5126, DOI:10.30591/jpit.v11i2.10186

Identifikasi Faktor Determinan Keterlibatan Pemain Game Online Berbasis Perilaku Menggunakan Machine Learning

Raid Alvaro Fathin Lie [1], Sindhu Rakasiwi [2]
[1,2]Universitas Dian Nuswantoro, Jl.Imam Bonjol, Pendrikan Kidul, 50131, Indonesia
[1]raidvvaro@gmail.com, [2]sindhu.rakasiwi@dsn.dinus.ac.id

Info Artikel

Riwayat Artikel:

Received 2026-01-14
Revised 2026-07-30
Accepted 2026-08-06

Abstract – *Along with the rapid growth of the online gaming industry and the increasing complexity of player behavioral patterns, player engagement analysis has become a critical component in designing data-driven retention strategies. This study aims to identify the determinant factors influencing player engagement levels in online games and to develop an accurate classification prediction model. The study is motivated by the limitations of conventional engagement analysis approaches that rely solely on total playtime, which tend to be biased and do not fully represent player loyalty. A quantitative research approach was employed using a behavioral dataset consisting of 40,034 players, with a comparative evaluation of Naïve Bayes, Logistic Regression, and Random Forest algorithms. The data preprocessing stage included variable encoding, feature scaling, and data partitioning using a stratified train–test split to preserve class distribution. Hyperparameter optimization for the Random Forest model was performed using Grid Search with 5-fold cross-validation to objectively determine the optimal parameter combination. Experimental results demonstrate that the Random Forest algorithm achieved the best performance with a test accuracy of 91.30%, outperforming Naïve Bayes (83.61%) and Logistic Regression (82.64%). Feature importance analysis revealed that demographic factors and total playtime contributed relatively little to player engagement prediction, whereas weekly gaming session frequency and average session duration were identified as the most influential determinants. The findings suggest that player retention strategies are more effective when shifting from a duration-based approach to a frequency-based approach, encouraging consistent player interaction and fostering long-term engagement.*

Keywords: *Behavioral Analysis; Data Mining; Online Games; Player Engagement; Random Forest.*

Corresponding Author:

Raid Alvaro Fathin Lie

Email: raidvvaro@gmail.com

Abstrak – *Seiring dengan pesatnya perkembangan industri game online dan meningkatnya kompleksitas pola perilaku pemain, analisis keterlibatan pemain menjadi aspek krusial dalam perancangan strategi retensi berbasis data. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor determinan yang mempengaruhi tingkat keterlibatan pemain dalam game online serta membangun model prediksi klasifikasi yang akurat. Penelitian ini dilatarbelakangi oleh keterbatasan pendekatan analisis keterlibatan yang hanya mengandalkan total durasi bermain, yang cenderung bias dan tidak sepenuhnya merepresentasikan loyalitas pemain. Metode penelitian menerapkan pendekatan kuantitatif menggunakan dataset perilaku sebanyak 40.034 pemain dengan membandingkan kinerja algoritma Naive Bayes, Logistic Regression, dan Random Forest. Tahap pra-pemrosesan data meliputi proses encoding variabel, penskalaan fitur, serta pembagian data menggunakan stratified train–test split untuk menjaga proporsi kelas. Optimasi hyperparameter pada algoritma Random Forest dilakukan menggunakan Grid Search dengan validasi silang 5-fold (5-fold cross-validation) untuk memperoleh kombinasi parameter terbaik secara objektif. Hasil eksperimen menunjukkan bahwa algoritma Random Forest memberikan kinerja terbaik dengan tingkat akurasi sebesar 91,30%, mengungguli algoritma Naive Bayes (83,61%) dan Logistic Regression (82,64%) sebagai model pembanding. Analisis feature importance mengungkapkan bahwa faktor demografis dan total jam bermain memiliki kontribusi yang relatif rendah, sementara frekuensi sesi bermain mingguan dan rata-rata durasi per sesi merupakan faktor determinan yang paling dominan. Penelitian ini menyimpulkan bahwa strategi retensi pemain lebih efektif apabila beralih dari pendekatan berbasis durasi ke pendekatan berbasis frekuensi untuk mendorong konsistensi interaksi dan mempertahankan keterlibatan pemain dalam jangka panjang.*

Kata Kunci*: Analisis Perilaku; Data Mining; Game Online*; Keterlibatan Pemain*; Random Forest.*

I. PENDAHULUAN

Dalam era digital, industri game online berkembang menjadi salah satu sektor ekonomi digital terbesar dengan model bisnis yang tidak lagi bergantung pada penjualan awal permainan, melainkan pada keberlangsungan keterlibatan pemain (*player engagement*) melalui pembelian dalam game, langganan, serta layanan berkelanjutan. Oleh karena itu, kemampuan untuk memprediksi tingkat keterlibatan pemain menjadi kebutuhan penting bagi pengembang dalam merancang strategi retensi pemain yang lebih efektif dan berbasis data [1], [2]. Seiring meningkatnya jumlah pemain, volume data perilaku yang dihasilkan selama aktivitas bermain juga semakin besar sehingga analisis manual menjadi kurang efisien. Kondisi tersebut mendorong pemanfaatan teknik *Machine Learning*

untuk menganalisis pola perilaku pemain secara lebih akurat dibandingkan pendekatan statistik konvensional [1], [2], [3].

Seiring meningkatnya jumlah pemain, volume data perilaku yang dihasilkan selama aktivitas bermain juga semakin besar sehingga analisis secara manual menjadi kurang efisien dan sulit mengidentifikasi pola perilaku yang kompleks [4], [5], [6], [7]. Berbagai penelitian memanfaatkan teknik Data Mining dan Machine Learning untuk menganalisis perilaku pemain berdasarkan aktivitas bermain, frekuensi sesi, durasi permainan, maupun riwayat interaksi pengguna [8], Pendekatan tersebut memungkinkan pengembang memahami karakteristik pemain secara lebih akurat dibandingkan analisis berbasis statistik sederhana.

Berbagai penelitian telah mengembangkan model prediksi perilaku pemain menggunakan beragam algoritma Machine Learning [9]. Penelitian oleh Reddy dkk. [1] membandingkan Random Forest, Support Vector Machine, Gradient Boosting, dan Long Short-Term Memory (LSTM) untuk memprediksi perilaku pembelian pemain berdasarkan aktivitas bermain. Sementara itu, Deloitte dkk. [2] membandingkan delapan algoritma Machine Learning pada kasus prediksi player churn dan menunjukkan bahwa Random Forest memberikan performa terbaik dengan nilai AUC lebih dari 0,99 dan akurasi mencapai 97% pada dataset permainan *The Settlers Online*. Penelitian yang lebih baru oleh Yuan [7] juga menunjukkan bahwa Random Forest mampu mengungguli Logistic Regression dalam klasifikasi popularitas game dengan akurasi sebesar 66,24%. Selain itu, Hartatik [10] berhasil memperoleh akurasi sebesar 90,8% menggunakan XGBoost untuk memprediksi tingkat keterlibatan pemain berdasarkan karakteristik demografis dan perilaku bermain, sedangkan Salsabila dkk. [11] melaporkan bahwa Artificial Neural Network memperoleh akurasi 90,76% dan mengungguli Logistic Regression yang hanya mencapai sekitar 82% pada dataset Kaggle yang sama.

Meskipun berbagai penelitian telah berhasil menghasilkan model prediksi dengan performa yang baik, sebagian besar penelitian berfokus pada peningkatan akurasi model melalui penggunaan algoritma yang semakin kompleks, seperti Artificial Neural Network maupun XGBoost [12], [13]. Penelitian-penelitian tersebut relatif belum banyak membahas bagaimana optimasi algoritma Random Forest menggunakan prosedur pencarian hyperparameter yang sistematis, sekaligus mengevaluasi faktor-faktor perilaku pemain yang paling berpengaruh terhadap tingkat keterlibatan. Selain itu, analisis kritis mengenai kontribusi masing-masing fitur terhadap pembentukan label target juga masih terbatas sehingga interpretabilitas model belum banyak dikaji.

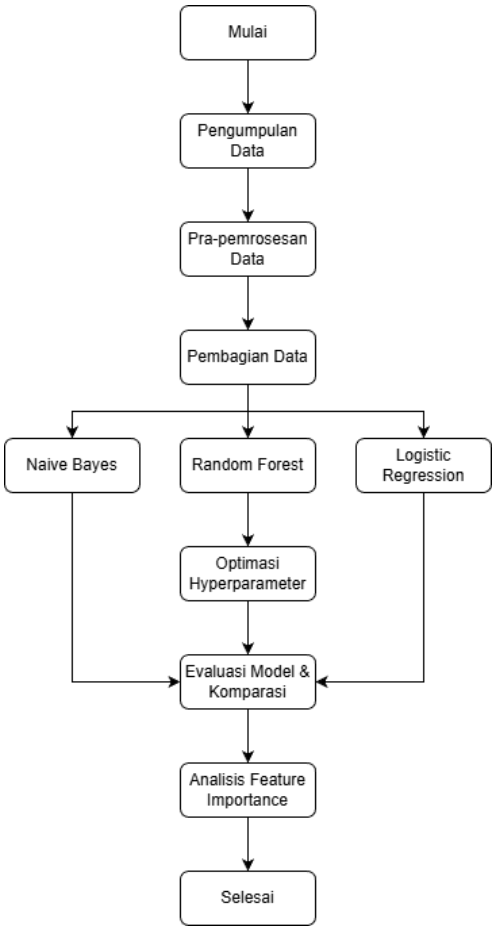
Berdasarkan research gap tersebut, penelitian ini mengusulkan model prediksi tingkat keterlibatan pemain menggunakan algoritma Random Forest yang dioptimalkan melalui GridSearchCV sebagai metode pencarian hyperparameter secara sistematis [14], [15], [16], [17], [18]. Berbeda dengan penelitian sebelumnya yang lebih berfokus pada peningkatan performa menggunakan algoritma yang lebih kompleks, penelitian ini menekankan keseimbangan antara akurasi dan interpretabilitas model melalui optimasi hyperparameter, evaluasi ROC-AUC multiclass, serta analisis feature importance untuk mengidentifikasi faktor-faktor dominan yang memengaruhi keterlibatan pemain [10], [19].

Kontribusi utama penelitian ini adalah: (1) membangun model prediksi tingkat keterlibatan pemain menggunakan Random Forest yang dioptimalkan melalui GridSearchCV; (2) membandingkan performa Random Forest terhadap Logistic Regression dan Naive Bayes menggunakan berbagai metrik evaluasi, meliputi accuracy, precision, recall, F1-score, confusion matrix, dan ROC-AUC multiclass; serta (3) mengidentifikasi faktor-faktor perilaku pemain yang paling berpengaruh melalui analisis feature importance.

Berdasarkan kontribusi tersebut, penelitian ini bertujuan untuk: (1) membangun model prediksi yang mampu mengklasifikasikan tingkat keterlibatan pemain ke dalam kategori Low, Medium, dan High; (2) mengevaluasi kinerja Random Forest hasil optimasi dibandingkan dengan Logistic Regression dan Naive Bayes; serta (3) menganalisis faktor-faktor dominan yang memengaruhi keterlibatan pemain sehingga dapat menjadi dasar penyusunan strategi retensi pemain berbasis data.

II. METODE

Alur penelitian ini dirancang dengan sistematis untuk memastikan validitas model prediksi. Tahapan penelitian dimulai dari pengumpulan data hingga evaluasi performa model, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Alur Metodologi Penelitian

Gambar 1 mengilustrasikan alur kerja sistematis penelitian ini, dimulai dari akuisisi data mentah yang dilanjutkan dengan tahapan pra-pemrosesan untuk menstandarisasi format data. Setelah data dibagi menjadi himpunan latih dan uji, proses pengembangan model dilakukan menggunakan tiga algoritma secara paralel (Naive Bayes, Logistic Regression, dan Random Forest). Secara khusus, algoritma Random Forest dioptimalkan menggunakan GridSearchCV yang secara sistematis mengevaluasi berbagai kombinasi hyperparameter pada data latih. Proses ini bertujuan memperoleh konfigurasi model terbaik berdasarkan hasil validasi tanpa melibatkan data uji, sehingga independensi proses evaluasi tetap terjaga sebelum dilakukan perbandingan kinerja model dan analisis feature importance [19].

Untuk meminimalkan potensi data leakage, dataset dibagi menjadi data latih dan data uji menggunakan teknik Stratified Train-Test Split sehingga proporsi setiap kelas target tetap terjaga pada kedua subset. Selanjutnya, proses standardisasi fitur menggunakan StandardScaler dilakukan setelah pembagian data, di mana parameter penskalaan dihitung hanya berdasarkan data latih dan kemudian diterapkan pada data uji. Tahapan optimasi hyperparameter menggunakan GridSearchCV juga dilakukan secara eksklusif pada data latih sehingga informasi dari data uji tidak memengaruhi proses pemilihan model. Seluruh fitur yang digunakan merupakan atribut perilaku pemain dan tidak mencakup variabel yang secara langsung merepresentasikan kelas target. Evaluasi akhir dilakukan menggunakan data uji melalui metrik Accuracy, Precision, Recall, F1-Score, Confusion Matrix, serta ROC-AUC multikelas untuk mengukur kemampuan generalisasi model dalam mengklasifikasikan tingkat keterlibatan pemain berdasarkan pola perilaku yang diamati.

Seluruh eksperimen dilakukan menggunakan Google Colaboratory dengan Python 3.12.13 dan pustaka Scikit-learn 1.6.1, Pandas 2.2.2, NumPy 2.0.2, Matplotlib 3.10.0, serta Seaborn 0.13.2 pada perangkat berprosesor Intel Core i5 11400H, GPU Nvidia Geforce RTX 3050, dan RAM 16 GB..

A. Pengumpulan Data

Penelitian ini memanfaatkan dataset publik *Online Gaming Behavior Insight* yang tersedia pada repositori Kaggle dan terdiri atas 40.034 data perilaku pemain game online [20]. Sebelum digunakan dalam proses pemodelan, dilakukan tahap pemeriksaan kualitas data (*data quality assessment*) untuk memastikan kelayakan dataset. Pemeriksaan meliputi verifikasi jumlah observasi dan atribut, identifikasi nilai hilang (*missing values*), pemeriksaan data duplikat, konsistensi tipe data setiap atribut, serta distribusi kelas pada variabel target *EngagementLevel*. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung nilai hilang maupun data duplikat sehingga seluruh sampel dapat digunakan dalam proses eksperimen.

Struktur dataset terdiri atas 12 variabel independen (*features*) dan satu variabel dependen (*target*) bernama *EngagementLevel* yang merepresentasikan tingkat keterlibatan pemain ke dalam tiga kategori, yaitu Low, Medium, dan High. Dokumentasi dataset tidak menjelaskan secara rinci mekanisme pembentukan label *EngagementLevel*. Oleh karena itu, penelitian ini menggunakan label sebagaimana disediakan oleh penyusun dataset tanpa melakukan rekonstruksi maupun modifikasi terhadap proses pelabelannya. Deskripsi karakteristik setiap fitur disajikan pada Tabel 1.

TABEL 1
DESKRIPSI FITUR DATASET

Fitur	Tipe Data	Deskripsi
PlayerID	Integer	Identitas unik pemain (dihapus saat pemrosesan)
Age	Integer	Usia pemain
Gender	Nominal	Jenis kelamin pemain
Location	Nominal	Lokasi geografis pemain
GameGenre	Nominal	Genre permainan (*RPG, Strategy, dll.)*
PlayTimeHours	Float	Total jam bermain
SessionsPerWeek	Integer	Jumlah sesi bermain per minggu
AvgSessionDuration	Integer	Rata-rata durasi per satu kali main (menit)
InGamePurchases	Integer	Status melakukan pembelian dalam *game* (*0/1*)
GameDifficulty	Ordinal	Tingkat kesulitan yang dipilih (*Easy/Med/Hard*)
PlayerLevel	Integer	Level karakter pemain saat ini
AchievementsUnlocked	Integer	Jumlah pencapaian yang dibuka
EngagementLevel	Ordinal	Target Klasifikasi (*Low,Medium,High*)

B. Pra-pemrosesan Data

Tahap pra-pemrosesan data bertujuan untuk memastikan bahwa data yang digunakan pada proses pembelajaran mesin memiliki format yang konsisten dan sesuai untuk setiap algoritma klasifikasi. Proses pra-pemrosesan meliputi penghapusan atribut yang tidak relevan, transformasi variabel kategorikal, pembagian data menjadi data latih dan data uji, serta standardisasi fitur numerik. Seluruh tahapan yang memanfaatkan informasi statistik data dilakukan hanya pada data latih untuk mencegah terjadinya *data leakage*.

Pada tahap ini, atribut PlayerID dihapus karena hanya berfungsi sebagai identitas unik pemain dan tidak memiliki kontribusi prediktif terhadap variabel target. Selanjutnya, variabel kategorikal diproses menggunakan dua pendekatan. Label Encoding diterapkan pada variabel ordinal, yaitu GameDifficulty (Easy = 0, Medium = 1, Hard = 2) serta variabel target EngagementLevel, sedangkan One-Hot Encoding digunakan pada variabel nominal seperti Gender, Location, dan GameGenre untuk menghindari terbentuknya hubungan ordinal semu antar kategori.

Setelah proses encoding selesai, dataset dipisahkan menjadi data latih dan data uji menggunakan Stratified Train-Test Split dengan rasio 80:20 sehingga proporsi setiap kelas tetap terjaga pada kedua subset. Selanjutnya dilakukan standardisasi fitur numerik menggunakan StandardScaler dengan metode normalisasi Z-score. Parameter standardisasi diperoleh hanya dari data latih (*fit*) dan kemudian diterapkan pada data uji (*transform*) guna mencegah terjadinya *data leakage*. Standardisasi ini terutama diperlukan pada algoritma Logistic Regression dan Gaussian Naive Bayes yang sensitif terhadap perbedaan skala fitur. Meskipun Random Forest secara teoritis tidak memerlukan standardisasi, seluruh model menggunakan data hasil pra-pemrosesan yang sama agar proses komparasi dilakukan pada kondisi eksperimen yang identik.

C. Pembagian Data

Dataset dibagi menggunakan teknik *Stratified Sampling* dengan rasio 80:20, di mana 80% data (32.027 sampel) digunakan untuk pelatihan (*training*) dan 20% data (8.007 sampel) digunakan untuk pengujian (*testing*). Teknik stratifikasi digunakan untuk menjaga proporsi sebaran kelas *Low*, *Medium*, dan *High* pada data latih dan uji tetap seimbang sesuai distribusi aslinya.

D. Model Development

Penelitian ini mengembangkan dan membandingkan tiga arsitektur model pembelajaran mesin untuk menangani masalah klasifikasi multikelas pada keterlibatan pemain. Fokus utama penelitian terletak pada pengembangan model *Random Forest* yang dioptimasi, dengan *Naive Bayes* dan *Logistic Regression* berfungsi sebagai model *baseline* (pembanding).

1. Random Forest Classifier Teroptimasi

Random Forest dipilih sebagai metode utama karena kemampuannya menangani noise dan mengurangi risiko overfitting yang umum terjadi pada Decision Tree tunggal melalui mekanisme Bootstrap Aggregating (Bagging). Algoritma ini membangun sekumpulan pohon keputusan dari subset data hasil proses bootstrap, kemudian menghasilkan prediksi akhir berdasarkan mekanisme majority voting antar pohon keputusan [21]. Untuk mengukur kualitas pemisahan (*split*) pada setiap simpul pohon, penelitian ini menggunakan Gini Impurity, yang dirumuskan sebagai berikut:

$$\text{Gini} = 1 - \sum_{i=1}^{C} (p_i)^2 \qquad (1)$$

Di mana P_i adalah probabilitas kelas ke- i dalam suatu simpul. Optimasi hyperparameter dilakukan menggunakan GridSearchCV dengan 5-fold cross-validation, sehingga setiap kombinasi parameter dievaluasi secara sistematis pada data latih untuk memperoleh konfigurasi model dengan performa validasi terbaik. Parameter yang dioptimasi meliputi jumlah pohon (n_estimators), kedalaman maksimum pohon (max_depth), dan jumlah minimum sampel untuk melakukan pemisahan simpul (min_samples_split). Pendekatan ini bertujuan meningkatkan kemampuan generalisasi model sekaligus mengurangi potensi overfitting sebelum dilakukan evaluasi pada data uji [22], [23].

2. Gaussian Naive Bayes

Sebagai pembanding probabilistik, penelitian ini menerapkan algoritma *Gaussian Naive Bayes*. Metode ini mengasumsikan bahwa fitur-fitur bersifat independen satu sama lain (*conditional independence*) dan nilai fitur kontinu terdistribusi secara normal (Gaussian). Probabilitas posterior untuk setiap kelas dihitung menggunakan Teorema Bayes:

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)} \qquad (2)$$

Meskipun asumsi independensi sering kali terlalu kuat untuk data dunia nyata, model ini memberikan *baseline* efisiensi komputasi yang tinggi dan performa yang stabil pada dataset berdimensi tinggi.

3. Multinomial Logistic Regression

Model pembanding kedua adalah *Logistic Regression* yang diperluas untuk klasifikasi multikelas (*Multinomial*). Berbeda dengan *Random Forest* yang non-linier, model ini memodelkan hubungan linier antara variabel independen dengan probabilitas logaritma (*log-odds*) dari variabel target. Fungsi *Softmax* digunakan untuk mengubah keluaran model menjadi distribusi probabilitas melintasi tiga kelas keterlibatan (*Low, Medium, High*):

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad (3)$$

Penggunaan model ini bertujuan untuk menguji hipotesis apakah pola keterlibatan pemain dapat dipisahkan secara linier (*linearly separable*) atau memerlukan batas keputusan yang kompleks seperti yang ditawarkan oleh *Random Forest*.

E. Evaluasi Kinerja Model

Untuk mengukur efektivitas model klasifikasi, penelitian ini menggunakan parameter evaluasi berbasis *Confusion Matrix*, yang meliputi *Accuracy*, *Precision*, *Recall*, dan *F1-Score*. Mengingat dataset memiliki distribusi kelas yang tidak seimbang (*imbalanced*), penggunaan *Accuracy* saja tidak cukup karena dapat memberikan bias pada kelas mayoritas. Oleh karena itu, *F1-Score* digunakan sebagai metrik harmonis yang menyeimbangkan *Precision* dan *Recall*, yang dirumuskan sebagai:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Sumbu Y: *True Positive Rate* (TPR) = $TPR \frac{TP}{TP + FN}$ (5)

Sumbu X: *False Positive Rate* (FPR) = $FPR \frac{FP}{FP + TN}$ (6)

Nilai $AUC = \int_0^1 TPR(FPR)d(FPR)$ (7)

Keterangan:

- TP (*True Positive*): Jumlah sampel dari kelas target yang diprediksi secara benar oleh model
- TN (*True Negative*): Jumlah sampel dari kelas lainnya (bukan target) yang diprediksi secara benar sebagai bukan bagian dari kelas target tersebut.
- FP (*False Positive*): Jumlah sampel negatif yang salah diprediksi sebagai positif (Kesalahan Tipe I).
- FN (*False Negative*): Jumlah sampel positif yang salah diprediksi sebagai negatif (Kesalahan Tipe II).

Kurva ROC memvisualisasikan *trade-off* antara *True Positive Rate* (sensitivitas) dan *False Positive Rate* pada berbagai ambang batas keputusan (*threshold*). Nilai AUC yang mendekati 1.0 mengindikasikan kemampuan diskriminasi model yang sangat baik dalam memisahkan antar kelas keterlibatan.

III. HASIL DAN PEMBAHASAN

Bab ini menyajikan temuan empiris dari eksperimen klasifikasi tingkat keterlibatan pemain *game online*. Pembahasan diawali dengan komparasi performa statistik antara model *Random Forest*, *Naive Bayes*, dan *Logistic Regression* untuk menentukan model terbaik. Analisis dilanjutkan dengan validasi keandalan model melalui kurva ROC-AUC, serta diakhiri dengan interpretasi mendalam mengenai faktor determinan perilaku (*Feature Importance*) yang paling berpengaruh terhadap retensi pemain.

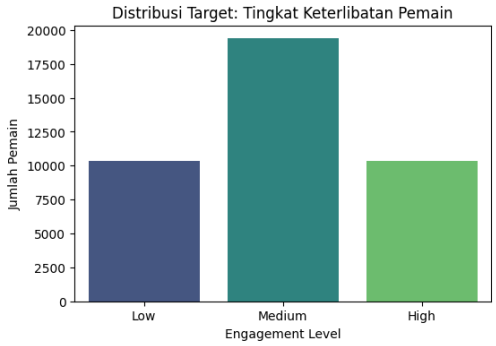
A. Analisis Statistik Deskriptif dan Eksplorasi Data

Sebelum dilakukan pemodelan prediktif, analisis statistik deskriptif dan eksplorasi data (*Exploratory Data Analysis/EDA*) dilakukan untuk memahami karakteristik dataset, distribusi setiap variabel, serta hubungan antarfitur. Tahap ini bertujuan mengidentifikasi karakteristik data yang dapat memengaruhi proses pemodelan sekaligus menentukan strategi pra-pemrosesan yang sesuai.

1. Analisis Keseimbangan Kelas Target

Analisis Keseimbangan Kelas Target, Langkah awal eksplorasi difokuskan pada distribusi variabel target EngagementLevel yang divisualisasikan pada Gambar 2. Hasil visualisasi menunjukkan bahwa distribusi kelas tidak sepenuhnya seimbang. Kelas Medium merupakan kelas dengan proporsi terbesar, yaitu sekitar 48,39% dari keseluruhan data, sedangkan kelas Low dan High masing-masing memiliki proporsi sebesar 25,79% dan 25,82%.

Ketidakseimbangan ini mengindikasikan bahwa mayoritas pemain dalam populasi dataset memiliki tingkat keterlibatan menengah. Kondisi ini menjadi landasan metodologis mengapa penelitian ini menerapkan teknik *Stratified Sampling* pada tahap pembagian data latih dan uji. Tanpa stratifikasi, pembagian data secara acak berisiko menghilangkan representasi kelas minoritas (*Low* atau *High*) dalam data latih, yang pada akhirnya dapat menyebabkan model menjadi bias dan cenderung memprediksi kelas mayoritas (*Medium*) saja. Oleh karena itu, pemahaman terhadap distribusi ini memvalidasi urgensi penanganan distribusi data sebelum proses pelatihan model dilakukan.



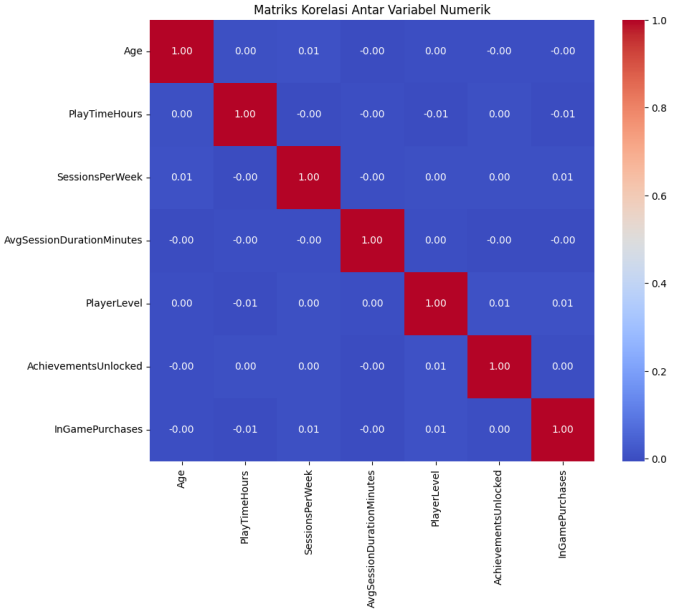
Gambar 2. Distribusi Kelas Target: Tingkat Keterlibatan Pemain.

2. Analisis Korelasi Antar Fitur

Analisis Korelasi Antar Fitur Selanjutnya, analisis korelasi *Pearson* dilakukan untuk menguji kekuatan hubungan linier antar variabel independen numerik, seperti PlayTimeHours, SessionsPerWeek, dan Age. Hasil matriks korelasi ditampilkan dalam bentuk *Heatmap* pada Gambar 3. Warna biru dan merah pada peta panas merepresentasikan kekuatan korelasi, di mana nilai mendekati 1 atau -1 menunjukkan hubungan linier yang kuat, sedangkan nilai mendekati 0 menunjukkan tidak adanya hubungan linier.

Temuan menarik dari visualisasi ini adalah mayoritas variabel menunjukkan koefisien korelasi yang sangat rendah (mendekati 0.00). Sebagai contoh, korelasi antara Age dan PlayTimeHours hampir bernilai nol, menunjukkan bahwa durasi bermain tidak memiliki pola linier terhadap usia pemain. Begitu pula antar variabel perilaku lainnya yang tidak menunjukkan multikolinearitas yang tinggi.

Absennya korelasi linier yang kuat ini memberikan wawasan teknis yang penting mengenai pemilihan algoritma. Algoritma berbasis linier seperti *Logistic Regression* bekerja dengan asumsi adanya hubungan garis lurus antar variabel. Dengan data yang memiliki korelasi linier sangat lemah seperti yang terlihat pada Gambar 3, wajar jika *Logistic Regression* kesulitan membentuk *decision boundary* yang optimal. Sebaliknya, hal ini memperkuat justifikasi penggunaan *Random Forest* sebagai model utama, karena algoritma ini bekerja berbasis aturan keputusan (*decision rules*) non-linier yang tidak bergantung pada asumsi korelasi linier antar variabel.



Gambar 3. Matriks Korelasi Antar Variabel Numerik.

B. Hasil Optimasi Hyperparameter

Sebelum dilakukan komparasi terhadap model klasifikasi, algoritma Random Forest terlebih dahulu melalui proses optimasi hyperparameter untuk memperoleh konfigurasi model yang memberikan performa terbaik. Optimasi dilakukan menggunakan GridSearchCV dengan 5-fold cross-validation, sehingga setiap kombinasi parameter dievaluasi secara sistematis berdasarkan rata-rata akurasi hasil validasi silang pada data latih. Parameter yang dioptimasi meliputi jumlah pohon keputusan (*n_estimators*), kedalaman maksimum pohon

(*max_depth*), dan jumlah minimum sampel untuk melakukan pemisahan simpul (*min_samples_split*). Ringkasan hasil optimasi disajikan pada Tabel 2. Berdasarkan hasil pencarian, konfigurasi parameter terbaik diperoleh pada kombinasi n_estimators = 200, max_depth = 20, dan min_samples_split = 2, yang menghasilkan nilai rata-rata akurasi validasi silang tertinggi. Konfigurasi tersebut kemudian digunakan sebagai model akhir (*best estimator*) untuk proses pengujian pada data uji. Konfigurasi tersebut kemudian digunakan sebagai **model akhir (best estimator)** pada tahap pengujian menggunakan data uji. Penggunaan GridSearchCV memastikan bahwa pemilihan hyperparameter dilakukan secara objektif melalui evaluasi berulang menggunakan validasi silang pada data latih sehingga mengurangi potensi bias pemilihan parameter. Selain itu, data uji tidak digunakan selama proses optimasi, melainkan hanya digunakan sekali pada tahap evaluasi akhir untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

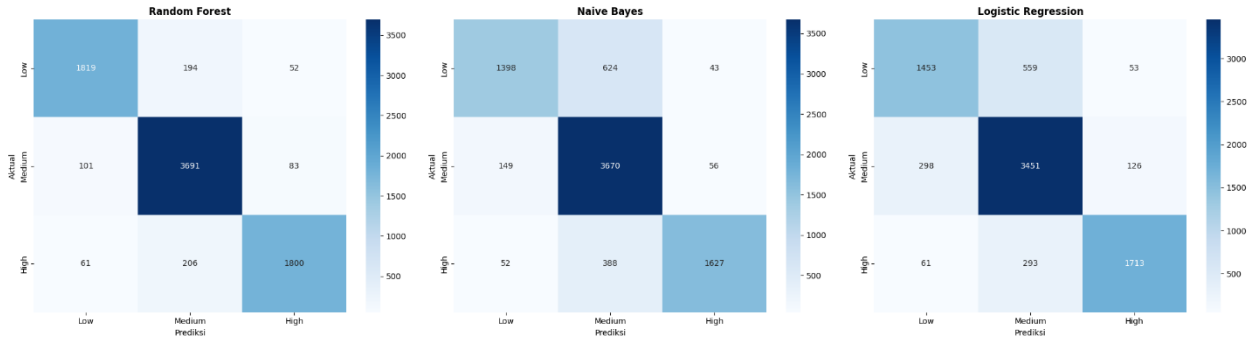
TABEL 2
HASIL OPTIMASI HYPERPARAMETER RANDOM FOREST

Parameter	Nilai Terbaik
n_estimators	200
max_depth	20
min_samples_split	2
Cross-validation	5-fold
Mean CV Accuracy	90.9764

C. Evaluasi Kinerja Klasifikasi

Eksperimen komparasi dilakukan terhadap tiga algoritma: *Random Forest*, *Naive Bayes*, dan *Logistic Regression* menggunakan data uji sebanyak 8.007 sampel. Hasil evaluasi model divisualisasikan melalui *Confusion Matrix* pada Gambar 4, yang memetakan perbandingan antara kelas aktual dan kelas prediksi. Secara keseluruhan, model *Random Forest* teroptimasi menunjukkan kinerja terbaik dengan akurasi sebesar 91,30%, lebih tinggi dibandingkan Gaussian Naive Bayes yang memperoleh akurasi (83,61%) serta Logistic Regression sebesar (82,64%). Dominasi *Random Forest* terlihat dari densitas warna biru tua yang sangat pekat dan konsisten pada garis diagonal matriks (*True Positive*), yang mengindikasikan tingginya ketepatan klasifikasi pada ketiga kelas (*Low, Medium, High*).

Analisis lebih lanjut menunjukkan bahwa Random Forest mampu mengklasifikasikan 1.819 sampel kelas Low, 3.691 sampel kelas Medium, dan 1.800 sampel kelas High secara benar. Kesalahan klasifikasi pada Random Forest relatif kecil, misalnya hanya 194 sampel kelas Low yang diprediksi sebagai Medium, serta 206 sampel kelas High yang diprediksi sebagai Medium. Pola kesalahan serupa juga ditemukan pada *Logistic Regression* yang salah memprediksi 559 sampel kelas *Low*. Sebaliknya, Naive Bayes masih mengalami kesulitan dalam membedakan kelas Low dari kelas Medium, yang terlihat dari 624 sampel kelas Low yang salah diprediksi sebagai Medium. Pola serupa juga ditemukan pada Logistic Regression, di mana 559 sampel kelas Low diprediksi sebagai Medium. Kesalahan klasifikasi tersebut menyebabkan kedua model memiliki akurasi yang lebih rendah dibandingkan Random Forest, terutama dalam mengenali pemain dengan tingkat keterlibatan rendah.



Gambar 4. Perbandingan *Confusion Matrix* antara *Random Forest*, *Naive Bayes*, dan *Logistic Regression*.

Untuk memberikan gambaran yang lebih komprehensif mengenai kualitas prediksi model pada setiap kelas, rincian metrik *Precision*, *Recall*, dan *F1-Score* disajikan pada Tabel 3. Berdasarkan tabel tersebut, terlihat adanya kesenjangan performa yang signifikan dalam menangani kelas minoritas (*Low* dan *High*).

Algoritma Naive Bayes dan Logistic Regression menunjukkan kelemahan pada metrik Recall untuk kelas Low, dengan nilai masing-masing hanya 0,68 dan 0,70. Angka ini mengindikasikan bahwa kedua model tersebut sering gagal mendeteksi pemain yang sebenarnya memiliki keterlibatan rendah (*False Negative*). Sebaliknya, Random Forest berhasil mempertahankan nilai Recall yang lebih tinggi, yaitu 0,88 untuk kelas Low dan 0,95 untuk kelas Medium. Stabilitas ini tercermin pada nilai *Weighted F1-Score* Random Forest yang mencapai 0,91, jauh mengungguli Naive Bayes (0,84) dan Logistic Regression (0,83). Temuan ini sejalan dengan penelitian Agustia dan Suryono [24] serta

Anggara dkk. [25], yang juga melaporkan bahwa Random Forest memberikan performa klasifikasi yang lebih baik dibandingkan Naive Bayes maupun Logistic Regression pada tugas klasifikasi. Hal ini menegaskan bahwa Random Forest merupakan model yang paling andal untuk diterapkan dalam skenario nyata yang memerlukan kemampuan deteksi pemain dengan tingkat keterlibatan rendah secara lebih akurat..

TABEL 3
PERBANDINGAN METRIK EVALUASI PER KELAS

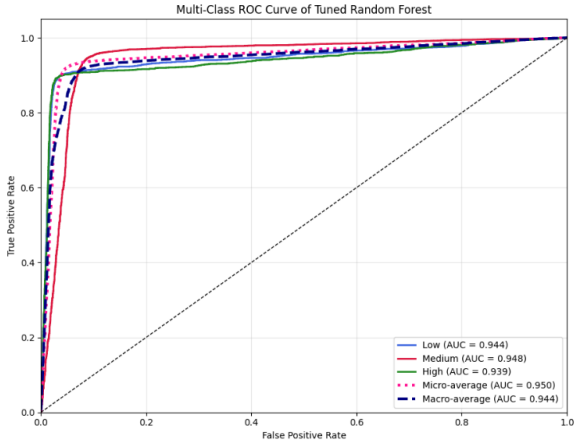
Model	Kelas	Precision	Recall	F1-Score
Random Forest (Accuracy: 0.91)	Low	0.91	0.88	0.89
	Medium	0.90	0.95	0.92
	High	0.93	0.87	0.89
Naive Bayes (Accuracy: 0.84)	Low	0.87	0.68	0.76
	Medium	0.78	0.95	0.86
	High	0.94	0.79	0.86
Logistic Regression (Accuracy: 0.83)	Low	0.80	0.70	0.75
	Medium	0.80	0.89	0.84
	High	0.91	0.83	0.87

D. Analisis Validasi Model (ROC-AUC)

Selain menggunakan metrik akurasi, validasi kemampuan diskriminasi model juga dilakukan melalui analisis Receiver Operating Characteristic (ROC) dan Area Under the Curve (AUC). Kurva ROC pada Gambar 5 mendeskripsikan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang (threshold), sehingga dapat menunjukkan kemampuan model dalam membedakan setiap kelas keterlibatan pemain.

Berdasarkan Gambar 5, model Random Forest yang telah dioptimasi menghasilkan nilai AUC sebesar 0.944 untuk kelas Low, 0.948 untuk kelas Medium, dan 0.939 untuk kelas High. Selain itu, diperoleh micro-average AUC sebesar 0.950 dan macro-average AUC sebesar 0.944. Seluruh nilai AUC berada di atas 0.90, yang menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik dalam membedakan ketiga tingkat keterlibatan pemain. Secara visual, seluruh kurva ROC berada jauh di atas garis diagonal yang merepresentasikan klasifikasi acak dan cenderung mendekati sudut kiri atas grafik. Pola tersebut menunjukkan bahwa model mampu mencapai True Positive Rate yang tinggi dengan False Positive Rate yang relatif rendah pada berbagai nilai threshold. Nilai AUC tertinggi diperoleh pada kelas Medium (0.948), sedangkan kelas High memiliki nilai AUC sedikit lebih rendah (0.939). Meskipun demikian, perbedaan tersebut relatif kecil sehingga menunjukkan bahwa kemampuan klasifikasi model tetap konsisten pada seluruh kelas.

Hasil analisis ROC-AUC ini memperkuat temuan pada evaluasi menggunakan confusion matrix dan metrik klasifikasi sebelumnya. Dengan nilai micro-average AUC sebesar 0.950 serta macro-average AUC sebesar 0.944, Random Forest tidak hanya mampu menghasilkan akurasi yang tinggi, tetapi juga memiliki kemampuan generalisasi yang baik dalam membedakan tingkat keterlibatan pemain berdasarkan pola perilaku yang diamati.



Gambar 5. Kurva *Multi-Class* ROC pada Model *Random Forest*.

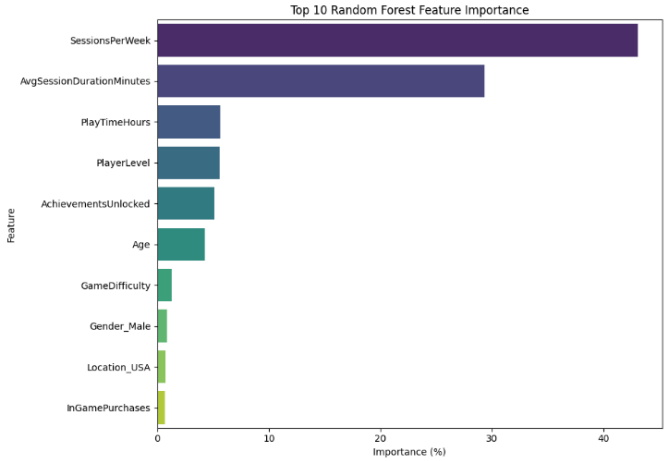
E. Evaluasi Klasifikasi Berdasarkan Confusion Matrix

Untuk memenuhi tujuan penelitian terkait faktor-faktor yang paling berpengaruh terhadap tingkat keterlibatan pemain, dilakukan analisis pentingnya fitur dengan menggunakan algoritma Random Forest berdasarkan penurunan nilai Gini Impurity. Hasil ranking fitur ditunjukkan pada Gambar 6. Hasil analisis menunjukkan bahwa SessionsPerWeek adalah fitur yang memberikan kontribusi terbesar dalam klasifikasi dengan nilai feature importance

sebesar 43,12%, sementara AvgSessionDurationMinutes menyusul di angka 29,35%. Secara total, kedua fitur itu berkontribusi sekitar 72,46% terhadap keseluruhan proses pengambilan keputusan model. Temuan ini menunjukkan bahwa frekuensi bermain dan rata-rata durasi setiap sesi merupakan indikator utama dalam menentukan tingkat keterlibatan pemain.

Sementara itu, PlayTimeHours hanya memiliki nilai feature importance sebesar 5,65%, diikuti oleh PlayerLevel (5,59%) dan AchievementsUnlocked (5,11%). Hasil ini mengindikasikan bahwa total akumulasi waktu bermain memiliki kontribusi yang lebih kecil dibandingkan pola bermain yang konsisten. Dengan kata lain, model lebih banyak memanfaatkan informasi mengenai seberapa sering pemain bermain dan lamanya setiap sesi dibandingkan hanya mempertimbangkan total waktu bermain.

Fitur-fitur lainnya, seperti Age (4,23%), GameDifficulty (1,30%), Gender, Location, GameGenre, serta InGamePurchases, memiliki kontribusi yang relatif rendah terhadap proses klasifikasi. Oleh karena itu, pada dataset yang digunakan, pola aktivitas bermain memberikan informasi yang jauh lebih dominan dibandingkan karakteristik demografis maupun atribut permainan dalam memprediksi tingkat keterlibatan pemain. Temuan ini konsisten dengan hasil evaluasi model sebelumnya, di mana Random Forest memperoleh akurasi sebesar 91,30% serta nilai macro-average AUC sebesar 0,944. Dominasi fitur SessionsPerWeek dan AvgSessionDurationMinutes menunjukkan bahwa keterlibatan pemain lebih dipengaruhi oleh konsistensi aktivitas bermain daripada karakteristik personal maupun total akumulasi waktu bermain



Gambar 6. Tingkat Kepentingan Fitur dalam Menentukan Keterlibatan Pemain.

F. Ablation Study – Pengaruh Fitur Terhadap Engagement level

Untuk memvalidasi kontribusi fitur-fitur utama terhadap performa model, penelitian ini melakukan ablation study dengan menghilangkan fitur yang memiliki nilai feature importance tertinggi secara bertahap, kemudian melatih kembali model Random Forest menggunakan prosedur optimasi hyperparameter yang sama seperti pada model dasar (baseline). Pada setiap skenario, proses optimasi dilakukan menggunakan GridSearchCV dengan 5-fold cross-validation, sehingga konfigurasi parameter terbaik diperoleh kembali sesuai dengan kombinasi fitur yang digunakan. Ringkasan hasil eksperimen disajikan pada Tabel 4.

TABEL 4
HASIL ABLATION STUDY PADA MODEL RANDOM FOREST

Eksperimen	Mean CV Accuracy (%)	Test Accuracy (%)
Baseline (Semua Fitur)	90.88	91.30
Tanpa *SessionsPerWeek*	57.40	57.32
Tanpa *AvgSessionDurationMinutes*	62.92	63.01
Tanpa Kedua Fitur	48.36	48.33

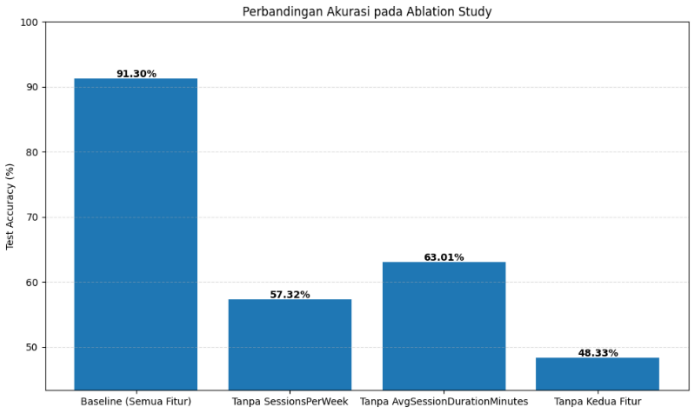
Berdasarkan Tabel 4, model dasar yang menggunakan seluruh fitur memperoleh mean cross-validation accuracy sebesar 90,88% dan akurasi data uji sebesar 91,30%. Ketika fitur SessionsPerWeek dihilangkan, akurasi data uji turun drastis menjadi 57,32%, atau mengalami penurunan sekitar 33,98 poin persentase dibandingkan model dasar. Penurunan ini merupakan yang terbesar di antara seluruh skenario, sehingga menunjukkan bahwa frekuensi bermain setiap minggu merupakan prediktor yang paling berpengaruh dalam menentukan tingkat keterlibatan pemain.

Eksperimen kedua menunjukkan bahwa penghapusan fitur AvgSessionDurationMinutes juga menyebabkan penurunan performa yang signifikan. Akurasi data uji turun menjadi 63,01%, atau berkurang sekitar 28,29 poin persentase dari model dasar. Hasil ini menunjukkan bahwa rata-rata durasi bermain pada setiap sesi memberikan informasi yang sangat penting bagi model dalam membedakan tingkat keterlibatan pemain, meskipun kontribusinya

masih berada di bawah SessionsPerWeek. Penurunan performa terbesar terjadi ketika kedua fitur tersebut dihilangkan secara bersamaan. Pada skenario ini, model hanya mampu mencapai 48,33% akurasi pada data uji dengan mean cross-validation accuracy sebesar 48,36%. Nilai tersebut mendekati performa klasifikasi acak untuk permasalahan tiga kelas, sehingga menunjukkan bahwa sebagian besar kemampuan prediksi Random Forest berasal dari kombinasi kedua fitur perilaku tersebut.

Hasil ablation study ini konsisten dengan analisis feature importance yang menunjukkan bahwa SessionsPerWeek dan AvgSessionDurationMinutes merupakan dua fitur dengan kontribusi terbesar terhadap proses klasifikasi. Dengan demikian, tidak hanya berdasarkan nilai feature importance, tetapi juga berdasarkan pengujian eksperimental melalui penghilangan fitur, kedua variabel tersebut terbukti merupakan faktor yang paling menentukan dalam memprediksi tingkat keterlibatan pemain pada dataset yang digunakan. Temuan ini memperkuat kesimpulan bahwa pola bermain yang konsisten, yang tercermin dari frekuensi bermain dan durasi setiap sesi, memiliki pengaruh yang jauh lebih besar dibandingkan karakteristik demografis maupun atribut permainan lainnya.

Implikasi praktis dari temuan ini menunjukkan bahwa strategi retensi pemain sebaiknya lebih berfokus pada peningkatan frekuensi interaksi pemain dibandingkan hanya mendorong akumulasi durasi bermain. Pengembang game dapat memanfaatkan informasi tersebut dalam merancang mekanisme permainan, seperti sistem daily login, misi harian, atau reward berbasis konsistensi bermain, sehingga mampu meningkatkan keterlibatan pemain secara berkelanjutan. Selain itu, model prediksi yang dihasilkan berpotensi digunakan sebagai alat pendukung keputusan untuk mengidentifikasi tingkat keterlibatan pemain secara lebih dini sehingga strategi retensi dapat diterapkan secara lebih tepat sasaran.



Gambar 7. Visualisasi Sampel Pohon Keputusan dari Model *Random Forest*.

G. Analisis Komparatif dengan Penelitian Terdahulu

Hasil penelitian ini menunjukkan bahwa algoritma Random Forest memberikan performa terbaik dibandingkan Gaussian Naive Bayes dan Multinomial Logistic Regression, dengan akurasi data uji sebesar 91,30%, serta nilai macro-average ROC-AUC sebesar 0,944. Temuan ini sejalan dengan berbagai penelitian terdahulu yang menunjukkan bahwa Random Forest merupakan salah satu algoritma yang memiliki kemampuan tinggi dalam menangani data perilaku pemain game yang bersifat kompleks dan nonlinier.

Penelitian oleh [2] mengenai prediksi churn pada permainan *The Settlers Online* melaporkan bahwa Random Forest menghasilkan performa terbaik dibandingkan beberapa algoritma klasifikasi lainnya dengan nilai AUC melebihi 0,99 dan akurasi mencapai 97%. Perbedaan performa tersebut dapat dipengaruhi oleh perbedaan karakteristik dataset, tujuan klasifikasi, serta proses rekayasa fitur (feature engineering) yang digunakan. Penelitian tersebut memanfaatkan data aktivitas permainan secara longitudinal dengan pendekatan *sliding window*, sedangkan penelitian ini menggunakan dataset publik yang bersifat potret (*cross-sectional*) untuk mengklasifikasikan tingkat keterlibatan pemain.

Hasil penelitian ini juga konsisten dengan penelitian [10] yang menerapkan algoritma XGBoost untuk memprediksi tingkat keterlibatan pemain dan memperoleh akurasi sebesar 90,8%. Meskipun menggunakan algoritma yang berbeda, kedua penelitian menunjukkan bahwa variabel yang berkaitan dengan intensitas aktivitas bermain, seperti frekuensi bermain dan durasi sesi, merupakan faktor yang paling berpengaruh terhadap tingkat keterlibatan pemain. Temuan tersebut memperkuat bahwa pola perilaku bermain memiliki kontribusi yang lebih besar dibandingkan karakteristik demografis pemain.

Selain itu, penelitian [11] yang membandingkan Artificial Neural Network (ANN) dengan Logistic Regression memperoleh akurasi 90,76% pada ANN dan sekitar 82% pada Logistic Regression. Hasil Logistic Regression pada penelitian ini menunjukkan pola yang serupa dengan akurasi sebesar 82,64%, sehingga mengindikasikan bahwa model linier memiliki keterbatasan dalam menangkap hubungan yang kompleks pada data perilaku pemain. Sebaliknya, Random Forest mampu memodelkan hubungan nonlinier antar fitur sehingga menghasilkan performa yang lebih baik.

Secara keseluruhan, hasil penelitian ini memperkuat temuan-temuan sebelumnya bahwa metode berbasis *ensemble learning*, khususnya Random Forest, merupakan pendekatan yang efektif untuk analisis perilaku pemain game. Selain menghasilkan performa klasifikasi yang tinggi, penelitian ini juga memberikan interpretasi mengenai faktor-faktor yang paling berpengaruh terhadap tingkat keterlibatan pemain melalui analisis feature importance dan didukung oleh hasil *ablation study*, sehingga tidak hanya berfokus pada peningkatan akurasi tetapi juga pada pemahaman perilaku pemain.

IV. SIMPULAN

Penelitian ini berhasil mengidentifikasi faktor-faktor yang mempengaruhi tingkat keterlibatan pemain game online serta membangun model klasifikasi menggunakan algoritma Naive Bayes, Logistic Regression, dan Random Forest. Berdasarkan hasil eksperimen, algoritma Random Forest yang dioptimasi menggunakan GridSearchCV dengan 5-fold cross-validation memberikan performa terbaik dengan akurasi data uji sebesar 91,30%, serta macro-average AUC = 0.944, sehingga mampu mengungguli Gaussian Naive Bayes (akurasi 83,61%) dan Logistic Regression (akurasi 82,64%). Analisis feature importance menunjukkan bahwa SessionsPerWeek dan AvgSessionDurationMinutes merupakan faktor yang paling berpengaruh terhadap tingkat keterlibatan pemain, sedangkan variabel seperti PlayTimeHours, usia, gender, lokasi, dan genre permainan memberikan kontribusi yang relatif lebih kecil. Temuan tersebut diperkuat melalui ablation study, yang memperlihatkan bahwa penghapusan kedua fitur utama menyebabkan penurunan akurasi model secara signifikan hingga 48,33%, sehingga mengonfirmasi bahwa kedua variabel tersebut merupakan komponen yang sangat penting dalam proses klasifikasi keterlibatan pemain. Hasil penelitian ini menunjukkan bahwa pola frekuensi dan konsistensi aktivitas bermain lebih representatif dalam memprediksi keterlibatan pemain dibandingkan hanya mengandalkan total durasi bermain. Implikasi praktis dari temuan ini adalah pengembang game dapat memanfaatkan model prediksi sebagai dasar dalam merancang strategi retensi yang lebih efektif, seperti fitur daily login, misi harian, maupun sistem penghargaan berbasis frekuensi bermain. Adapun untuk penelitian selanjutnya disarankan melakukan validasi pada dataset dari berbagai jenis permainan serta membandingkan Random Forest dengan algoritma klasifikasi lain untuk menguji generalisasi model.

DAFTAR PUSTAKA

[1] K. N. Reddy, G. Nikhil Kumar, K. Anitha, J. Mounika, and B. Prasad, "Player Behavior Prediction for In-Game Purchases Using Machine Learning," *J. Adv. Dev. Res. IJAIDR26011798*, vol. 17, no. 1, Accessed: Jul. 27, 2026. [Online]. Available: www.ijaidr.com•

[2] K. R. Deloitte, N. Pflanzl, J. A. Hüllmann, and M. Preuss, "Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game", Accessed: Jul. 27, 2026. [Online]. Available: www.thesettlersonline.com/.

[3] M. Ahmad and J. Srivastava, *Behavioral data mining and network analysis in massive online games*. 2014. doi: 10.1145/2556195.2556196.

[4] J. Kristensen and P. Burelli, *Combining Sequential and Aggregated Data for Churn Prediction in Casual Freemium Games*. 2019. doi: 10.1109/CIG.2019.8848106.

[5] H. Hoang and N. Cam, "Do they like your game? Early-stage churn prediction using a two-phase neural network system," *Eng. Appl. Artif. Intell.*, vol. 144, p. 110102, Mar. 2025, doi: 10.1016/j.engappai.2025.110102.

[6] J. Kim, K. Kim, H. Kim, and S. Li, "Analyzing player churn in esports games: a survival analysis of league of legends log data," *Sport. Bus. Manag. An Int. J.*, vol. 16, pp. 453–474, Jan. 2026, doi: 10.1108/SBM-04-2025-0093.

[7] L. Yuan, "Predicting Online Gaming Player Behavior and Engagement Based on Machine Learning," *Appl. Comput. Eng.*, vol. 223, pp. 54–62, Feb. 2026, doi: 10.54254/2755-2721/2026.AS31576.

[8] A. Rawat *et al.*, *Predictive User Engagement Modeling Using Time Series Analysis for Interactive Media and Gaming*. 2025. doi: 10.1109/GEM66882.2025.11155680.

[9] R. Aprillya, P. : Perbandingan, M. Klasifikasi, R. A. Putri, and N. S. Fatonah, "Perbandingan Metode Klasifikasi serta Analisis Faktor Akademis Pola Kelulusan Mahasiswa di Perguruan Tinggi," *J. Inform. J. Pengemb. IT*, vol. 7, no. 2, pp. 109–117, May 2022, doi: 10.30591/JPIT.V7I2.3082.

[10] H. Hartatik, "Safe Gaming Analytics: Gender- and Age-Aware Machine Learning Using XGBoost for Game Player Engagement Prediction," *West Sci. Inf. Syst. Technol.*, vol. 2, pp. 419–427, Dec. 2024, doi: 10.58812/wsist.v2i03.2150.

[11] R. A. Salsabila, A. Herita, N. Zahra, W. Sari, K. Tania, and F. Fathoni, "PREDIKSI TINGKAT KETERLIBATAN PEMAIN GAME ONLINE BERDASARKAN INTENSITAS DAN POLA AKTIVITAS BERMAIN MENGGUNAKAN PENDEKATAN HYBRID MACHINE LEARNING," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 10, pp. 4054–4061, May 2026, doi: 10.36040/jati.v10i3.18152.

[12] Y. Jung, H. Jang, and S. Kim, "Explainable AI for Player Retention: Hyperpersonalized Promotions to Prevent Game User Churn," *IEEE Trans. Games*, vol. 18, pp. 369–382, Jan. 2026, doi: 10.1109/TG.2026.3689469.

[13] E. Loria and A. Marconi, "Exploiting limited players' behavioral data to predict churn in gamification," *Electron. Commer. Res. Appl.*, vol. 47, p. 101057, May 2021, doi: 10.1016/j.elerap.2021.101057.

[14] H. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[15] M. K. Suryadi, R. Herteno, S. W. Saputro, M. R. Faisal, and R. A. Nugroho, "Comparative Study of Various Hyperparameter Tuning on Random Forest Classification With SMOTE and Feature Selection Using Genetic Algorithm in Software Defect Prediction," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 2, pp. 137–147, Mar. 2024, doi: 10.35882/JEEEMI.V6I2.375.

[16] F. Hasan, R. Khatun, D. Raza, H. H. Tuhin, and B. Lab, "Chapter 7 - Hyper Parameter Optimization Using Machine Learning," 2025, p. 43.

[17] A. Paul, D. Mukherjee, P. Das, A. Gangopadhyay, A. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Trans. Image Process.*, vol. 27, pp. 4012–4024, May 2018, doi: 10.1109/TIP.2018.2834830.

[18] P. Probst, A.-L. Boulesteix, and M. Wright, *Hyperparameters and Tuning Strategies for Random Forest*. 2018. doi: 10.48550/arXiv.1804.03515.
[19] S. Yuni, M. Oktavianingrum, and F. Aksa, "Analisis Feature Importance pada Penyakit Alzheimer Menggunakan Random Forest," *J. Data Sci. Methods Appl.*, vol. 2, pp. 49–57, Jun. 2026, doi: 10.30873/jodmapps.v2i1.1464.
[20] W. A. Yasir, *Online Gaming Behavior Insight*, Kaggle Dataset, 2025. doi:10.34740/kaggle/dsv/14228009.
[21] R. A. Prasetyo, W. T. Saputro, and D. Chirzah, "Perbandingan Logistic Regression, SVM, dan Random Forest untuk Analisis Sentimen Ulasan Aplikasi Gopay," *J. Inform. J. Pengemb. IT*, vol. 10, no. 4, pp. 1176–1188, Sep. 2025, doi: 10.30591/JPIT.V10I4.8796.
[22] K. Wisatawan Mancanegara Di Prov Sulawesi Selatan Dengan K-Means Dan SVM Nero Caesar Gosari, "Penerapan Data Mining Dalam Mengelompokkan Kunjungan Wisatawan Mancanegara Di Prov. Sulawesi Selatan Dengan K-Means Dan SVM," *J. Inform. J. Pengemb. IT*, vol. 8, no. 3, pp. 174–180, Sep. 2023, Accessed: Jul. 27, 2026. [Online]. Available: https://ejournal.poltekharber.ac.id/index.php/informatika/article/view/4554
[23] H. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.
[24] D. Agustia and R. Suryono, "Comparison of Naïve Bayes, Random Forest, and Logistic Regression Algorithms for Sentiment Analysis Online Gambling," *INOVTEK Polbeng - Seri Inform.*, vol. 10, pp. 284–295, Jan. 2025, doi: 10.35314/prk93630.
[25] K. Anggara, R. Gea, H. Upendar, and R. Fahlapi, "Perbandingan Naïve Bayes dan Random Forest untuk Klasifikasi Sentimen Ulasan Produk Amazon Fire HD 7," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 5, pp. 1369–1378, Jul. 2026, doi: 10.62712/juktisi.v5i2.1315.